Annotation Guidelines for Corpus Novelties: Part 2 - Alias Resolution

Arthur Amalvy, Vincent Labatut

Annotation Guidelines for Corpus *Novelties*: Part 2 – Alias Resolution

Version 1.0.0

30 September 2024

Arthur Amalvy & Vincent Labatut

Laboratoire Informatique d'Avignon – LIA UPR 4128, Avignon, France

Abstract

The *Novelties* corpus is a collection of novels (and parts of novels) annotated for Alias Resolution, among other tasks. This document describes the guidelines applied during the annotation process. It contains the instructions used by the annotators, as well as a number of examples retrieved from the annotated novels, and illustrating how canonical names should be defined, and which names should be considered as referring to the same entity.

Contents

Abstract 1

Contents 2

1 Introduction 3
- 1.1 Notion of *Alias* 3
- 1.2 Organization 3

2 Annotation Process 4
- 2.1 Preparation of the Resources . . . 4
- 2.2 Processing of an Entity Name . . . 4
- 2.3 Overall Verification 5
- 2.4 Metadata and Finalization 5

3 Canonical Forms 6
- 3.1 Character Entities (`CHR`) 6
- 3.2 Location Entities (`LOC`) 8
- 3.3 Organization Entities (`ORG`) 9
- 3.4 Group Entities (`GRP`) 9
- 3.5 Miscellaneous Entities (`MSC`) 10

4 Concluding Remarks 12

A Version History 13

References 14

1 Introduction

This document aims at providing instructions for the annotation of aliases in the *Novelties* corpus. The corpus itself will be the object of a separate description. It was constituted mainly to fulfill two goals: in the short term, train and test NLP methods able to handle *long* texts, and in the longer term, be used to develop *Renard* [2], a pipeline aiming at extracting *character networks* from literary fiction. This pipeline includes several processing steps besides alias resolution, including named entity recognition and coreference resolution. Character networks can be used to tackle a number of tasks, including the assessment of literary theories, the level of historicity of a narrative, detecting roles in stories, classifying novels, identify subplots, segment a storyline, summarize a story, design recommendation systems, align narratives, etc. See the detailed survey of Labatut and Bost [6] for more information regarding character networks. There are seldom annotation guidelines for alias resolution in the literature, so the one presented here are designed from scratch, taking into account this application's context.

1.1 Notion of *Alias*

In this document, aliases are the different names or expressions used in a novel to refer to the same entity. The goal of alias resolution is to automatically determine whether two expressions refer to the same entity. The task in not well-defined in the literature in general, and is sometimes termed differently. Alternative names include "character detection", "character identification" or "character clustering". While some works are specifically concerned with alias resolution [5, 7], most solve the task in the pursuit of another one, such as character classification [3] or speaker attribution [4]. Still, alias resolution is an important step of character network extraction, which is why we choose to include this annotation layer in our corpus.

In the context of *Novelties*, the manual annotation process consists in elaborating a unique name for each entity, called a canonical form, and associate it to each occurrence of the entity. When performing the network extraction, the vertices are named using these canonical forms.

In the rest of the document, we provide a number of examples to illustrate our guidelines. Text extracted from novels is formatted using a sans-serif font, e.g. the Emperor.

1.2 Organization

In the following, we first describe guidelines for the alias annotation process from a very operational perspective (Section 2). We then list the rules used to elaborate the canonical forms, depending on the type of the considered entities (Section 3). Finally, Section 4 provides our concluding remarks, and Appendix A gives the history of this document.

2 Annotation Process

At this stage, we assume that the considered novel has already been annotated for Named Entity Recognition, using our guidelines [1]. It is better if the person that annotated the named entities also performs the alias annotation, as they have more experience with the story and characters, and are less likely to make mistakes.

The goal of this section is to provide a detailed and very concrete description of the process we use, so that anyone can easily reproduce it. We do not use a specific tool, but rather a spreadsheet editor such as Calc or Excel, and possibly a simple text editor to correct errors in the `.conll` files. Concretely, the annotation process goes as follows.

2.1 Preparation of the Resources

The few operations described in the following can significantly help to perform the whole process, and to reduce the number of mistakes.

Entity and mention lists We use the scripts available on the Novelties GitHub repository to extract two CSV files:

1. Entity list: the list of all unique entity names, with their entity type and frequency;
2. Mention list: the list of all entity mentions, with their positions in the novel (chapter and line).

The fourth column of the first file is empty, and designed to receive the canonical form of each entity name. The rest of the process consists in filling this column.

For the sake of convenience, we order this file first by entity name, then by entity type. The former criterion allows grouping relatively similar names, and the latter allows processing the entity types separately.

Secondary resources The next step is to open all necessary secondary resources. These will be used mainly to search additional details, and to perform some verifications. These include:

- Some Web content such as a fan Wiki[1] dedicated to the novel, or some annotated version of the text[2]. This is useful to determine the canonical name of certain entities.
- The epub version of the novel, which can be viewed in a software such as Calibre[3]. This allows efficiently searching through the book to make various verifications.
- The list of entity mentions extracted at the previous step. It can be opened using a spreadsheet editor, but does not require any sorting, unlike the list of entity names. It can be convenient to filter the rows by entity type, though.

2.2 Processing of an Entity Name

For each entity name in the list, we perform the following operations.

Fill the canonical form In the CSV file (entity list), we add the canonical form associated to the current entity name. Very often, this amounts to pasting the same name. For instance, if one is processing Harry Potter, then one will use the exact same string as the canonical form.

Check suspect names If the current name looks non-canonical, or if it could be an alias in the strict sense (i.e. a name completely different of the character's real name), it is important to check in the novel and in the secondary sources. For instance, if one is processing Voldemor, then one needs to retrieve the canonical form Tom Marvolo Riddle from a secondary source. The rules to determine the canonical form are described in Section 3.

[1] e.g. A Wiki of Ice and Fire: https://awoiaf.westeros.org/index.php/Main_Page
[2] e.g. Power Moby Dick: http://www.powermobydick.com/
[3] https://calibre-ebook.com/

Search for other occurrences Once the canonical form is set, we look for parts of the entity name in the rest of the CSV file, in order to find other names referring to the same entity. For instance, if the current name can be split into a first name and a last name (ex. Harry Potter), one can look for both separately (Harry and Potter). If the other entity names already have a canonical form, it is important to verify that they are exactly the same as the current one. If they do not, then one can associate the current canonical form to these other entity names.

Correct annotation errors Sometimes, the current entity name looks incorrect: it is visibly not a character name (autobus), or it contains non-alphabetic characters (Potter!!!), or it seems incomplete (Potte), etc. In this case, there is probably an error in one of the `.conll` files containing the NER annotation. To correct this mistake, we use the mention list (i.e. the other CSV file) to precisely locate the error, and directly edit the `.conll` file. We also correct the problem in the entity list (but not in the mention list), for future verification.

2.3 Overall Verification

Once all the names in the entity list have been processed as described before, we get what we call a v1 of the alias annotations. It is now necessary to perform certain verifications in this list.

Unicity of the Canonical Forms First, we verify that the same canonical form is not written in several different ways (e.g. minor character case differences). In the now annotated entity list, generate the list of unique names. When using LibreOffice Calc:

1. Select the column containing canonical forms;
2. Go to menu *Data* > *More Filters* > *Standard Filter*
3. In the dialog box, *Filter Criteria*, set *Field Name* to `none`;
4. In *Options*, check *No duplications* and *Copy results to*, then select the target column;
5. Uncheck *Keep filter criteria* and *Case sensitive*

If a canonical form presents any variability, we correct this.

Regeneration of the Lists As in the first step (Section 2.1), we use the scripts to generate the entity list and mention list. While doing so, we must make sure to keep the v1 file intact, as we need both versions of the entity list. In the following, v2 refers to the updated entity list.

Verification of the Entity Annotations This operation is particularly important, especially if some corrections have been conducted in the `.conll` files during the annotation process. We copy the content of the v1 file and paste it in the v2 file, near the existing content. In an empty column, we now define a formula that checks whether the v1 canonical form is equal to the v2 canonical form. We paste the formula over all rows. Finally, we check whether all the values in this new column are `TRUE`. Any `FALSE` value reveals a difference between v1 and v2, generally corresponding to a modification incorrectly performed during the second step (Section 2.2). It is necessary to properly apply them, by performing the appropriate correction in the `.conll` files.

After these corrections, we must do the verification process again: regenerate the entity list, and compare it to v1. Only when both lists are exactly similar can we conclude the whole process.

2.4 Metadata and Finalization

Once we reach a stable annotation file, we create a new, fifth, column, entitled `metadata`, in order to store the following information, similarly to what we do in the NER annotation files [1]:

- Title of the novel (field `Title`);
- Name of the annotators (field `Annotator`);
- Version of the guidelines (field `Guidelines`);
- Date of last update (field `Updated`).

Finally, we rename the file as `alias_resolution.csv` and place it in the same folder as the NER annotation files.

3 Canonical Forms

We follow a certain number of rules to define the canonical form of an entity's name. The general idea is that it should be complete enough to clearly identify the entity unambiguously. Or at least, as much as possible depending on the novel and entity. These rules are slightly different from one entity type to the other. It is important to stress that the canonical form does not necessarily appear in the original text. For example, it can come from an external source.

3.1 Character Entities (`CHR`)

General Rule The general rule for character entities is to form the canonical name based only on the first and family names. See for instance Constance Bonacieux, in Alexandre Dumas's *The Three Musketeers*:

Form in the text	Canonical form
Constance	Constance Bonacieux
Constance Bonacieux	
Madame Bonacieux	
Mme. Bonacieux	

We ignore honorifics (here Madame and Mme.), unless the first name or the last name is unknown. In this case, we include the main honorific title, for instance:

Form in the text	Canonical form
Comte de Wardes	Comte de Wardes
De Wardes	
M. de Wardes	
Monsieur de Wardes	
Monsieur le Comte de Wardes	
Coquenard	Monsieur Coquenard
M. Coquenard	
Monsieur Coquenard	

Tripartite Names Certain cultures use names composed of three parts, in which case our canonical form includes all three of them. This is the case, in particular, of Latin (example from *The Three Musketeers*) and Russian (example from Dostoevsky's *The Double*) names:

Form in the text	Canonical form
Caesar	Gaius Julius Caesar
Yakov Petrovitch Golyadkin	Yakov Petrovitch Golyadkin
Yakov Petrovitch	
brother Yakov	
Mr. Golyadkin	

Other types of novels are likely to exhibits different types of names, e.g. completely made-up systems in Fantasy novels, or context-dependent names in the Chinese culture[4]. The general principle here is to adapt the nature of the canonical names to the considered novel.

Aliases If a character appears under completely different aliases, we try to use their actual name, e.g. for D'Artagnan's antagonist Milady in *The Three Musketeers*:

[4] https://en.wikipedia.org/wiki/Chinese_name#Alternative_names

Form in the text	Canonical form
Anne de Breuil	Anne de Breuil
Charlotte Backson	
Comtesse de la Fère	
Comtesse de Winter	
Lady Clarik	
Milady	
Milady Clarik	
Milady de Winter	

Nicknames When the character possesses a nickname that is important to its identification, it is included in the canonical form as a complement. For instance, in *The Three Musketeers*, the main character is only known as D'Artagnan, and one of the musketeers is simply Athos:

Form in the text	Canonical form
D'Artagnan	Charles de Batz de Castelmore, dit d'Artagnan
Lord d'Artagnan	
M. d'Artagnan	
Monsieur d'Artagnan	
Athos	Olivier de La Fère, dit Athos
Comte de la Fère	
Monsieur Athos	

As mentioned before, the canonical form does not necessarily appear in the original text: it may come from a secondary source, as is the case here.

Historical Characters Certain authors like to mention historical characters, to provide some context to their story. We retrieve the full name from secondary sources, even when they are not used in the novel. For instance, in *The Three Musketeers*:

Form in the text	Canonical form
Brutus	Marcus Junius Brutus
Cervantes	Miguel de Cervantes
Robespierre	Maximilien de Robespierre

Nobility Ranks We proceed similarly when the character possesses a nobility rank mainly used to identify them. Consider for example one of the main antagonists, Cardinal Richelieu:

Form in the text	Canonical form
Cardinal de Richelieu	Armand Jean du Plessis, duc de Richelieu
Cardinal Richelieu	
M. de Richelieu	
Monseigneur the Cardinal	
Monseigneur the Cardinal Richelieu	

Kings and Queens When dealing with a king or queen, we include the kingdom in the canonical form, in order to avoid any confusion with other monarchs with the same name. Take Francis the First, which is mentioned in *The Three Musketeers*:

Form in the text	Canonical form
Francis I	Francis I of France
Francis the First	

The novel refers to the king of France, and not to the Emperor of Austria bearing the same name.

God & Satan The Abrahamic god appears under a number of names, which we all associate to the canonic form God. In *The Three Musketeers*, for instance:

Form in the text	Canonical form
Dieu	God
Father	
Gad	
God	
Holy Father	
Holy Ghost	
Son	

Note that we also annotate the Catholic trinity as God, for the sake of simplicity.

We proceed similarly with Satan and all his names. For instance, in Herman Melville's *Moby Dick*:

Form in the text	Canonical form
Beelzebub	Satan
Devil	
Evil One	
Lucifer	
Prince of the Powers of the Air	
Satan	

3.2 Location Entities (LOC)

General Rule For locations, the general rule is straightforward, and consists in just using the proper noun as the canonical form:

Form in the text	Canonical form
America	America
Amiens	Amiens
Angers	Angers
Angoutin	Angoutin
Anjou	Anjou

Noun Modifiers However, sometimes the proper noun is ambiguous when considered separately, and one needs to add a modifier. For instance, in *The Three Musketeers*:

Form in the text	Canonical form
Abbey St. Germain	Abbey St. Germain
St. Germain	Faubourg St. Germain

Here, the first example refers to a monastery, whereas the second one is a suburb. Even when there is no ambiguity, it is better to keep a modifier if it helps to understand what the name refers to, e.g. Mediterranean sea, Mount Kilimanjaro. This makes it easier to identify what the entity is, even without needing to check its type (e.g. LOC).

Expressions using road are a bit specific. See these examples from *The Three Musketeers*:

Form in the text	Canonical form
road to Picardy	Picardy
road of Chaillot	Road of Chaillot

The context for the first example above is: "While his Eminence was seeking for me in Paris, I would take, without sound of drum or trumpet, the road to Picardy, and would go and make some inquiries concerning my three companions." We understand that road to Picardy refers to a transportation means rather than a place, and what is important here is the destination, i.e. Picardy. The context for the second example is: "Between six and seven o'clock the road of Chaillot is quite deserted; you might as well go and ride in the forest of Bondy." Here, the author specifically refers to the road of Chaillot as a location, where some event takes place.

Part of Locations Sometimes, the proper noun of a location is completed in order to refer to a part of this location. In this case, we consider the relevance of the smaller location in the story: is

it important to make the distinction with the larger location? If the answer is yes, then we keep the precision, otherwise, we use the same canonical form as for the larger entity. For instance, in *Moby Dick*:

Form in the text	Canonical form
Banks of Newfoundland	Newfoundland
coast of Greenland	Greenland

3.3 Organization Entities (ORG)

Political Entities In order to clearly distinguish organizations from locations, we explicitly indicate the nature of the organization in the canonical form, even if it is not specified in the novel. In *The Three Musketeers*:

Form in the text	Canonical form
Austria	Archduchy of Austria
Denmark	Kingdom of Denmark
France kingdom of France	Kingdom of France
Paris	City of Paris

Commercial Entities It is sometimes not clear that an entity is a commercial structure based on its name alone, so we generally include a precision in its canonical form. In *The Three Musketeers*:

Form in the text	Canonical form
Golden Lily	Inn of the Golden Lily
Jolly Miller	Hostel of the Jolly Miller
Post	Tavern of the Post

3.4 Group Entities (GRP)

Families In order to be very clear that a family name does not refer to a specific person but to the whole group, we explicitly add House to the canonical form. For instance, in *The Three Musketeers*:

Form in the text	Canonical form
Condés	House Condé
Montmorency	House of Montmorency

We use the word House because it is explicitly used in this novel. But in a more general case, one could use Family instead: the most important point is to be consistent over the whole novel.

Demonyms & Ethnonyms When dealing with group names that reflect a geographic or ethnical origin, we use the plural form. If no such form exists in English, then we add people to the adjective. See these examples from *The Three Musketeers*:

Form in the text	Canonical form
Andalusian	Andalusians
Arabian	Arabs
Assyrians	Assyrians
Béarnais Béarnese	Béarnese people
Berrichan Berrichon	Berrichons
English Englishman Englishmen Englishwoman	Englishmen

If there are several different forms distinguishing gender, we adopt the majority rule, and use the one with the most occurrences as the canonical form. In the above example, Englishman (36 occurrences)

and Englishmen (10) are more frequent than Englishwoman (5), which is why we use Englishmen as the canonical form.

3.5 Miscellaneous Entities (`MSC`)

Food When the entity refers to food, we generally specify it in the canonical form, especially if it can be confused with another type of entity (such as a location). In *The Three Musketeers*:

Form in the text	Canonical form
Bordeaux	Bordeaux wine
Burgundy	Burgundy wine
chambertin	Chambertin wine
champagne	Champagne wine

Bordeaux is a city, and Burgundy and Champagne are regions, in France.

Events and Periods The principle is the same when dealing with historical events or periods: we want to avoid any confusion. In *The Three Musketeers*:

Form in the text	Canonical form
Chalais	Henri de Talleyrand, marquis de Chalais
Chalais	Chalais's conspiracy

Here, the Conspiracy of Chalais was a plot against Richelieu that took place before the story told in *The Three Musketeers*. But Chalais is also the name of the most central person in this plot.

Regarding periods, we also complete the name that appears in the novel, if necessary. Here are some example from Glen Cook's *The Black Company*:

Form in the text	Canonical form
ancient kingdom	Ancient Kingdom period
Cho'n Delor	Cho'n Delor era

Works For intellectual and artistic works, we build the canonical form based on the author's name and the work's title. In *The Three Musketeers*:

Form in the text	Canonical form
Augustinus	C. Jansenius's *Augustinus*
Iliad	Homer's *Iliad*
L'Avare	Molière's *L'avare*

When the text refers to someone's corpus of work more globally, we use the name of the author and mention its whole work:

Form in the text	Canonical form
St. Augustine	St. Augustine's work
St. Bartholomew	Bartholomew's work
St. Chrysostom	St. Chrysostom's work

If the author is unknown, we simply indicate the nature of the work, like in these examples from Aldous Huxley's *Brave New World*:

Form in the text	Canonical form
Hug me till you drug me, honey	*Hug me till you drug me, honey* song
Three Weeks in a Helicopter	*Three Weeks in a Helicopter* movie

Scriptures We use Bible as the canonical form for all parts of the Bible. In *The Three Musketeers*, for instance:

Form in the text	Canonical form
Bible	Bible
Scripture	
Scriptures	
Judith	

Here, Judith refers to the *Book of Judith* in the *Old Testament*.

Languages We explicitly state the nature of a language in its canonical form, to avoid any confusion with demonyms. In *The Three Musketeers*:

Form in the text	Canonical form
English	English language
French	French language
German	German language

4 Concluding Remarks

The annotation of aliases is much more time-consuming than one would expect *a priori*. Indeed, it requires performing many verifications in order to make sure that we get a proper bijection between the sets of entities and surface forms. This is particularly true for novels taking place in historical settings (e.g. *The Three Musketeers*, *Eugénie Grandet*, *The Red and the Black*), as one must recover the full names of certain entities, that are not always mentioned in the book. Moreover, these novels mix fictional and real entities, that must also be checked against some reliable reference (e.g. Wikipedia, or a specialized Wiki). Novels such as *Moby Dick*, with many cultural references, are also hard: their annotation is significantly eased through the use of a commented version of the novel. Generative models can also be of some help, especially when dealing with historical references in classic novels, although at the time of writing, their answers must be carefully verified. We typically provide the model with an excerpt of the novel and ask it about the specific entity.

Another time-consuming aspect of this annotation tasks, is that it reveals a number of issues among the NER annotations. This mainly concerns the boundaries of the annotations (too many words, too few words, `B-xxx` instead of `I-xxx` or the opposite); their type (missing type, incorrect type); and inconsistencies in the NER annotation process throughout the novel (e.g. some definite description not annotated systematically). In any case, detecting such issues implies correcting the NER annotation, and performing a new verification of the aliases.

Determining whether two surface forms refer to the same entity can be quite easy when one form is just a variant of the other (e.g. D'Artagnan vs. M. D'Artagnan). But it can also be very tricky, when one entity has completely different names (cf. Milady in *The Three Musketeers*, Section 3.1). The latter case requires a relatively good understanding of the story. For this reason, it is much more convenient, easy, and efficient, that 1) the same person defines the NER and alias annotations; and 2) the alias annotation is conducted right after the NER annotation, when the story is still fresh in the mind of the annotator.

A Version History

We use three-part version numbers of the form major–minor–patch for both these guidelines and the *Novelties* corpus. See the named entity annotation guidelines for more details [1].

Version	Date	Changes	Corpus
`1.0.0`	30/09/2024	First version, based on the annotation of a few novels.	

Version `1.0.0` This is the first version of our guidelines for alias resolution. It is based on the annotation of a few books including fantasy (*The Blade Itself*, *The Black Company*), Science-Fiction (*1984*, *Brave New World*) and classic (*Moby Dick*, *The Three Musketeers*, *Eugénie Grandet*) novels.

References

[1] A. Amalvy and V. Labatut. Annotations guidelines for corpus *novelties*: Part 1 – named entity recognition. Technical report, Avignon Université, 2024.

[2] A. Amalvy, V. Labatut, and R. Dufour. Renard: A modular pipeline for extracting character networks from narrative texts. *Journal of Open Source Software*, 9(98):6574, 2024. doi:10.21105/joss.06574.

[3] D. Bamman, T. Underwood, and N. A. Smith. A bayesian mixed effects model of literary character. In *52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 370–379, 2014. doi:10.3115/v1/P14-1035.

[4] C. Cuesta-Lazaro, A. Prasad, and T. Wood. What does the sea say to the shore ? a BERT based DST style approach for speaker to dialogue attribution in novels. In *60th Annual Meeting of the Association for Computational Linguistics*, pages 5820–5829, 2022. doi:10.18653/v1/2022.acl-long.400.

[5] L. Jahan, R. Mittal, W. V. Yarlott, and M. Finlayson. A straightforward approach to narratologically grounded character identification. In *28th International Conference on Computational Linguistics*, pages 6089–6100, 2020. doi:10.18653/v1/2020.coling-main.536.

[6] V. Labatut and X. Bost. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys*, 52(5):89, 2019. doi:10.1145/3344548.

[7] H. Vala, D. Jurgens, A. Piper, and D. Ruths. Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Conference on Empirical Methods in Natural Language Processing*, pages 769–774, 2015. doi:10.18653/v1/D15-1088.